ARMY RESEARCH LABORATORY

Method to Select Technical Terms for Glossaries in Support of Joint Task Force Operations

by Michelle Vanni

ARL-TN-0467 **January 2012**

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

Army Research Laboratory
Adelphi, MD 20783-1197

ARL-TN-0467 **January 2012**

Method to Select Technical Terms for Glossaries in Support of Joint Task Force Operations

Michelle Vanni
Computational and Information Sciences Directorate, ARL

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED (From - To)
January 2012	Final	

4. TITLE AND SUBTITLE
Method to Select Technical Terms for Glossaries in Support of Joint Task Force Operations

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Michelle Vanni

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
U.S. Army Research Laboratory
ATTN: RDRL-CII-T
2800 Powder Mill Road
Adelphi, MD 20783-1197

8. PERFORMING ORGANIZATION REPORT NUMBER
ARL-TN-0467

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT
Our technique for selecting foreign-language technical terms for human-use glossaries and automatic processor lexicons offers a customized solution based on sound principles that has resulted in an effectiveness breakthrough for the Army. Prepared in support of a single Joint Task Force (JTF), its principle-based underpinnings justify its use in similar applications for various JTFs involved in strategic operations. Language induces expectations from its community of use that can be exploited to provide more effective machine translation (MT). Entries in finely tuned glossaries and lexicons, which are devoid of ambiguity, carry a valence that activates readers' associated world knowledge. The clarity of the entry builds reader confidence while the activated semantic fields permit readers a higher likelihood of accurate context interpretation than would otherwise be possible. A glossary prepared for one document in particular serves to illustrate the method employed throughout the project for glossary development. This document and a human translation are freely available at http://alep.stanford.edu/.

15. SUBJECT TERMS
MT, IE, text analytics

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Michelle Vanni
Unclassified	Unclassified	Unclassified	UU	16	19b. TELEPHONE NUMBER (*Include area code*) (301) 394-0367

Contents

INTENTIONALLY LEFT BLANK.

1. Introduction

The complexity of the Army mission often impels the scientist toward creative problem-solving. Principled approaches to Army challenges will draw pertinent features from several related disciplines. Missions are unique. So, techniques considered are usually as yet unimplemented in academic and industrial research and development (R&D) settings.

Nevertheless, with testing and enhancement in mission-oriented labs such as the U.S. Army Research Laboratory (ARL), customized solutions result in *effectiveness* breakthroughs. Moreover, when solutions are based on sound and relevant principles and disciplines, they also carry the potential for Army reuse to inform a general framework.

The technique outlined here, for selecting the technical terms to include in glossaries designed to aid human linguists with foreign language word and phrase look-up and enhance automatic processes, such as machine translation (MT), was developed in support of Combined Joint Inter-Agency Task Force 435 (CJIATF-435) in Afghanistan.

As detailed below, ARL had taken responsibility for a coordinated solution to address the need for rapid and high quality Afghan language translation with a proposal for integrated materiel development, which could subsequently be leveraged by an existing Army machine foreign language translation acquisition program such as the Machine Foreign Language Translation System (MFLTS), the Army's Program of Record for MT.

The work of a Joint Task Force (JTF) is, by definition, a team performance by individuals with varying expertise, perspectives, and skills, toiling together toward common goals. While the present method underpins a capability that serves only one specific group, the foundation of the method explained here justifies its use to inform glossary building and MT tasks for similar JTFs operating at various strategic locations.

2. Motivation and Background

We were doubly motivated in undertaking this particular task. Our primary interest was in meeting an immediate mission need and, in fulfilling that requirement, we also wanted to exploit a fundamental principle of linguistic cognition, that is, that language use induces expectations.

2.1 Mission Requirements

At the request of the Office of Director Defense Research and Engineering (DDR&E), the Army Director of Capabilities Integration, Prioritization, and Analysis drew up an Execution Plan to respond to the Joint Urgent Operational Need of Improved Machine Based Language Translation

for Afghan Languages (JUON CC-0429). Section Four of that plan states that "[ARL] will build a comprehensive glossary of organization names, acronyms, and technical terms from the legal and criminal justice domains with a target size of 5,000 words, […] will compile this new Task Force glossary as an electronic file using acronyms and other items found on the HarmonieWeb portal at the Rule of Law site, [and] will elicit glossary items from points of contact at the CJIATF headquarters […] and other staff sections." It goes on to indicate that English-Dari as a language pair will be given priority over English-Pashto and that the glossaries will be reformatted for dual use as user-specific dictionaries in MT software.

2.2 Expectation Grammar Theory

It goes without saying that one wants the words in one's language technology glossary to include those domain terms that occur in the material that supports the bilingual work. Thus, physicians want to see medical terms, attorneys want to see legal terms, and Soldiers want to see military terms. This is the assumption that, reasonably, underpins the concept of *user-dictionary* as a feature of the linguists' automated look-up tools.

Less frequently noted is the logic behind the incorporation of these valuable handcrafted resources—characterized by extremely precise renderings—into tools for the *automatic processing* of semantic equivalence of text in two or more languages, as in, for example, automatic translators and multilingual summarizers. Since language is a system of *signs*, or sound-meaning duals, agreed upon by a community, it is a human phenomenon that develops not only individually but also, necessarily, at group level. In fact, throughout life, our *idiolect*, or individual system of linguistic choices, becomes an important means by which we express our identity with one or more groups.

Viewed another way, the language use of the group or subgroup can constitute a sublanguage, often referred to as *jargon* for a professional group or *argot* for less well-defined groups. Meanings associated with jargon terms are tied to a specific concept, typical activity, or prevalent attitude displayed in that community. What makes them valuable for the purposes of automatic language processing, especially as embodied in finely tuned glossaries, is that their semantic structure is devoid of ambiguity.

For readers of MT output who are familiar with the jargon of the community served by the incorporated glossary, encountering and understanding a well-translated term with a specific sense, or a well-rendered name with a specific referent, is akin to the second language learners' experience of encountering, in a challenging second language text, a word or phrase that they understand the meaning of and actually realize that they understand it. In second language learning, this is known as "comprehensible input" (Krashen, 1981).

This "input" to the MT reader's cognitive process is helpful in two ways. First, it gives an unambiguous sense to the segment translated and thereby increases the reader's confidence in the fidelity of the automatic rendering as a whole. Second, it triggers world or "professional world"

knowledge, which is related to the irrevocably understood concept. The freshly triggered knowledge permits the reader to interpret the text with a higher probability of accuracy than would otherwise be possible. According to one theorist, with confidence, understanding and related knowledge, readers generate "grammar-based expectancies" or hypotheses about event sequences analogous to the plans of the author, in the MT condition, of the source language text (Oller, 1983).

The congruent expectancies then increase the likelihood of accurate understanding. This idea was supported in an ARL pilot study of human acceptability judgments on MT output. Acceptability judgments were compared on two versions of output, one in which names were accurately rendered (A Set) and another in which names were inaccurately rendered (I Set). Using a Magnitude Estimation (ME) methodology in which subjects made a direct numerical estimation of the degree to which sentences in the data conveyed the meaning in the reference sentences, investigators found a 34.8% difference. There was 22% difference between A Set and I Set scores, using automatic evaluation in Meteor (Lavie and Agarwal, 2007).* A differential effect was thus detected, suggesting that weighting proper name rendering in automated evaluation systems may improve the reliability of these systems.

3. Methodology

A glossary prepared for one document in particular, the first text in Stanford Law School's Afghan Legal Education Project (ALEP), *An Introduction to the Law of Afghanistan*, will serve to illustrate the method employed throughout the project for glossary development. This document and a human translation of it are freely available online (Stanford, 2011).

3.1 Automation

The first step is an automatic process for culling frequently occurring content words from a text. This step is necessary when a text is particularly lengthy. The ALEP document contained 234 pages, so, a human effort alone would have been prohibitively time-consuming. Instead, we identified two publicly available terminology extractor tools: TerMine (NaCTEM, 2011) and Alchemy (AlchemyAPI, 2011; Rose, 2011).

TerMine evaluates a candidate term based on four corpus statistical characteristics related to the term: its length, its occurrence frequency, its frequency as part of other longer candidate terms, and the number of these longer candidate terms. The formula that determines *termhood* and is incorporated into the algorithm is called *C-value*. This measure accounts for nested terms by

*Meteor is an automatic metric for MT evaluation, which has demonstrated high correlation with human judgments of translation quality, significantly outperforming the more commonly used Bleu metric.

recognizing term context words and then incorporating information from the context words into the term extraction process (Frantzi, Ananiadou, and Mima, 2000).

The AlchemyAPI Web site provided two tools, an Interactive Demonstration and a Keyword/Terminology Extractor. Output from the former, which constituted a subset of the latter, was marked by high precision, and that from the latter, by high recall. The Alchemy approach contrasts with that used in TerMine in that Alchemy will process the text with information categories, such as person, location, and organization, in addition to returning topic keywords. Output from both TerMine and Alchemy Keyword/Terminology Extractor were submitted for human-in-the-loop selection.

3.2 Selection and Context of Use

The criteria used in selecting terms for this project follow conceptual constructs in the *corpus linguistics* research literature, especially "context of use" (Biber, Conrad, and Reppen, 1998). According to this principle, a word or expression can have a unique meaning within a given community or situational setting. The word or multi-word expression can also be associated with that context without any change in general meaning.

When settings and groups determine an agreed-upon sense, an expression may occur outside a given context of use, but with a different meaning. For example, "lower house" and "upper house" in a "governing" context refer to legislative assemblies. The exact same phrase, however, in a "geographic location" context, refer to the placement of residences. This is not to say that the sense, let's call it S1, in the first context, which we'll call C1, can never occur in the second context, C2, and vice versa. It only indicates that S1 exists and is distinct from the sense it has in the second context, S2.

When a sense, S1, associated with a context, C1, does occur in a well-defined separate context, C2, there may still be subtle changes along different semantic lines of, for example, register or emphasis. Full names are examples of expressions in a formal register that refer to a single person, S1, and are usually reserved for formal occasions and documents, C1, such as ceremonies and forms. But in informal settings, C2, such as family gatherings, full names, while maintaining their S1 reference, may merely be used for emphasis with the referent, the person being referred to, remaining unchanged.

What is important about the "sense by association" aspect of the "context of use" principle is that a word or expression can also be highly correlated with a group or context while maintaining the same meaning both inside and outside of the context. The frequency and typicality of occurrence in one given context of use is what gives the word or expression its unique meaning and its membership in a semantic category associated with that context.† Consider the case of the conjunction, "notwithstanding," a word associated with legal and other formal written contexts.

†Its occurrence at the syntactic level is beyond the scope of this report.

Its sense as a marker of discourse function is not lost in other, non-legal, contexts, only less frequent and typical. As a result, "notwithstanding" might be considered part of the legal lexicon.

With this in mind, we populated the glossary for this project with words and expressions having a unique meaning or high frequency of occurrence in the specific context of use of nation-building in Afghanistan, as judged by examination and semantic analysis of the CJIATF-435 material.

The CJIATF-435 is tasked with setting standards of behavior for detention facilities, defining elements of parliamentary structure, reporting on police actions, and providing lessons for training, background for leadership development, and information for and about other initiatives. They thus rely on text types, such as press reports, presentations, handbooks, and instruction manuals, among other material. There is no one domain or one genre that adequately captures the linguistic variety that the resources under construction will be designed to handle.

Because the notion of context-of-use transcends the traditional concepts of domain and genre, it is a useful rubric for deciding which lexical items logically to include in the JUONS glossaries, customizers, and bilingual training datasets. Single lexical items, as well as multi-word noun-based and verb-based expressions will be found in the lists. Technical terminology, as a category of reference within communities of common interest (CCI), is a set of words whose context of use is kept constant. For example, among the medical community *in medical settings*, one hears the terms, *coronary infarction*, *arterial sclerosis*, *edema*, *angina*, etc. As a category then, technical terms in a CCI function in a manner similar to named entities in a CCI consisting of speakers acquainted with the named entity. That is, for each term, CCI and context-of- use, there is only one sense and, for each named-entity, CCI and context-of-use, there is only one referent.‡

Technical terms are generally included in the glossaries. They may also be embedded in context-of-use expressions. In these cases, the term is extracted to stand alone as a single lexical item. The rest of the expression is then reevaluated according to the criteria described earlier. As for the forms included, we limited ourselves, for this first pass, to the forms that occurred in the material, leaving the questions of which ideally to include or how ideally to process the forms for future work.

‡Again, the issue of ambiguous references, to include those to entity referents with the same name or one entity with two names—open research questions in their own right—is beyond the scope of this report.

3.3 Translation

The next step in the process of lexical development to support the building of mission-specific statistical machine translation (SMT) systems and glossaries is the translation of the selected terminology. Once a final list of terms is established, the developer inputs the selections to the latest SMT system-in-progress to produce a list of translated terms.[§]

For all the reasons that the SMT is still incomplete, that is, faulty alignments, out-of-domain training data, and inconsistent segmentation and spelling, among others, the list of translations in the output is sparse and error-laden. However, the ratio of the number of valid or fairly close renderings to the number of decidedly unhelpful ones is generally high enough to justify the effort in automation. The bilingual list output assists the project's native speaker linguists, or "humans-in-the-loop," by saving them time and tedium searching for and consulting about appropriate terminological translations. In this way, the selection step serves also to lighten the burden on the native speaker linguist whose job it is to ensure the quality of the SMT support to the mission project, not unlike the pipeline mechanism for preparing bilingual corpora for the linguist's review; see Tanenbaum, LaRocca, and Morgan (2011).

4. Discussion and Future Work

Progress in the direction of greater automation without a sacrifice of quality in this context relies on the vast linguistic knowledge that can only be supplied by the human language specialist, subject matter expert, and linguist. Thus, much of the automation developed and used in support of mission-focused MT development goes toward facilitating the work of the human linguist, that is, alleviating repetitive, tedious, and time-consuming tasks. For example, the pipeline system, noted in section 3, is geared to harvesting, cleaning, aligning, and presenting to a language specialist two semantically equivalent texts, in different languages, segment by segment.

Without that automation, the highly qualified language specialists would be obliged to spend much of their time cutting-and-pasting the texts from the Web page and reformatting it to eliminate noise elements, as encountered. This means their work would consist, for the most part, of deleting reoccurring mark-up, stray text, framing, and advertisements; editing misspellings, spacing, and formatting errors; and inserting appropriate text for both halves of the bilingual corpus. Needless to say, pipeline automation affords the project a considerable savings in terms of the cost and mental fatigue of the linguists/language specialists, who, with the

[§]Morgan (2011) describes the project-specific, human-in-the-loop SMT system-building methodology.

pipeline, can stay energized by contributing their unique and sophisticated linguistic acumen to the effort.

The same is true of the term selection process. Burdening the linguist and subject matter expert with the task of repeatedly translating frequently occurring terms and substantives, which are errorful and light on content, makes inefficient and inappropriate use of their time and talents, which, at the end of the day, is cost ineffective. By contrast, what we have presented here is a method for term selection that is based on sound and relevant principles of automation and linguistics. Its value lies in its mission effectiveness, which can only be measured by putting it into practice for human-in-the-loop foreign language system and resource development. If, with use, the latter serves to increase Soldier effectiveness, then it is our hope that the method will become a standard and that the concepts it embodies will inform a general framework for development of foreign language glossaries and MT resources.

5. References

AlchemyAPI Web site. http://www.alchemyapi.com/api/keyword/ (accessed November 2011).

Biber, Douglas; Conrad, Susan; Reppen, Randi. *Corpus Linguistics: Investigating Language Structure and Use*; Cambridge University Press, 1998.

Department of the Army Memorandum. Office, Director of Defense Research and Engineering (DDR&E). 29 Nov 2010. Joint Urgent Operational Need for Improved Machine-Based Language Translation for Afghan Languages (JUON CC-0429).

Frantzi, K.; Ananiadou, S.; Mima, H. Automatic Recognition of Multi-word Terms. *International Journal of Digital Libraries* **2000**, *3* (2), 117–132.

Krashen, Stephen D. *Principles and Practice in Second Language Acquisition*; English Language Teaching series, London: Prentice-Hall International (UK) Ltd, 1981.

Lavie, A.; Agarwal, A. 2007. Meteor: an Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. *Proceedings of Second Workshop on Statistical MT, StatMT '07*. Stroudsburg, PA: Association for Computational Linguistics, 228–231.

Morgan, John J. *Project-specific Machine Translation*; ARL-TR-????; U.S. Army Research Laboratory: Adelphi, MD, 2011.

NaCTEM Web site. http://www.nactem.ac.uk/software/termine/ (accessed November 2011).

Oller, John W.; Richard-Amato, Patricia A. Eds. *Methods that Work*; Rowley, MA: Newbury House, (pp 4–5), 1983.

Rose, Gabriella. *Assessment of Publicly-available Resources to Build an In-house Term Extractor*; not published yet; U.S. Army Research Laboratory: Adelphi, MD, 2011.

Stanford Law School Web site. Afghanistan Legal Education Project. http://alep.stanford.edu/ (accessed November 2011).

Tanenbaum, William; LaRocca, Steve; Morgan, John. *Introduction of Automation for the Production of Bilingual Aligned Parallel Text*; ARL-TR-5798; U.S. Army Research Laboratory: Adelphi, MD, 2011.

Vanni, M.; Walrath, J. *Differential Effect of Correct Name Translation on Human and Automated Judgments of Translation Acceptability: A Pilot Study*; ARL-TR-4630; U.S. Army Research Laboratory: Adelphi, MD, September 2008.

List of Symbols, Abbreviations, and Acronyms

ALEP	Afghan Legal Education Project
ARL	U.S. Army Research Laboratory
CCI	communities of common interest
CJIATF-435	Combined Joint Inter-Agency Task Force 435
DDR&E	Office of Director Defense Research and Engineering
JTF	Joint Task Force
ME	Magnitude Estimation
MT	machine translation
R&D	research and development
MFLTS	Machine Foreign Language Translation System
SMT	statistical machine translation

NO. OF COPIES	ORGANIZATION
1 PDF	ADMNSTR DEFNS TECHL INFO CTR ATTN DTIC OCP 8725 JOHN J KINGMAN RD STE 0944 FT BELVOIR VA 22060-6218
1	DARPA ATTN IXO S WELBY 3701 N FAIRFAX DR ARLINGTON VA 22203-1714
1 CD	OFC OF THE SECY OF DEFNS ATTN ODDRE (R&AT) THE PENTAGON WASHINGTON DC 20301-3080
1	US ARMY RSRCH DEV & ENGRG CMND ARMAMENT RSRCH DEV & ENGRG CTR ARMAMENT ENGRG & TECH CTR ATTN AMSRD AAR AEF T J MATTS BLDG 305 APG MD 21005-5001
1	US ARMY INFO SYS ENGRG CMND ATTN AMSEL IS TD A RIVERA FT HUACHUCA AZ 85613-5300
1	COMMANDER US ARMY RDECOM ATTN AMSRD AMR W C MCCORKLE 5400 FOWLER RD REDSTONE ARSENAL AL 35898-5000

NO. OF COPIES	ORGANIZATION
1	DIRECTOR US ARMY RSRCH LAB ATTN AMSRD ARL RO EV W D BACH PO BOX 12211 RESEARCH TRIANGLE PARK NC 27709
1	DARPA J OLIVE 3701 N FAIRFAX DR ARLINGTON VA 22203-1714
17 HCS 1 PDF	US ARMY RSRCH LAB ATTN RDRL CII B BROOME ATTN RDRL CII T M T VANNI (12 HCs, 1 PDF) ATTN RDRL CII T M HOLLAND ATTN RDRL CIO MT TECHL PUB ATTN RDRL CIO LL TECHL LIB ATTN IMNE ALC HRR MAIL & RECORDS MGMT ADELPHI MD 20783-1197

TOTAL: 26 (23 HCs, 2 PDFs, 1 CD)